Binary Text Classification Across Levels of Expertise: Statistical and Contextual Approaches Using Zipf, USE and BERT

November 14, 2024

Abstract

This paper investigates three independent approaches for text classification using data collected from both the English and Simple English Wikipedia across several text categories. First, we explore a statistical method based on the Zipf distribution, specifically Cumulative Distribution Function (CDF) derived from the Zipf curve, to classify text. This approach leverages the statistical properties of word frequency distributions to derive features for classification. Second, we employ the Universal Sentence Encoder (USE) to generate semantic embeddings and compute cosine similarity between test text embeddings and the centroid of each class's text embeddings. This method focuses on capturing contextual meaning and classifies texts based on their proximity to pre-defined class representations. Finally, we apply a fine-tuned BERT-based classifier to categorize texts as either beginner-level or expert-level, utilizing BERT's deep contextual understanding of language to classify texts based on their complexity. Each method is evaluated independently, and the results demonstrate that while Zipf provides a statistical lens on text classification, USE and BERT bring richer, context-aware insights, offering a comprehensive approach to classifying text at different levels of abstraction.

Keywords: Binary Text Classification, Zipf's Law, Statistical Methods, Contextual Embeddings, Universal Sentence Encoder (USE), BERT, Natural Language Processing (NLP), Semantic Similarity, Cosine Similarity, Expert vs. Beginner Text Classification, Word Frequency Distribution, Transformer

Models, Embeddings, Text Complexity, Machine Learning for Text, Linguistic Features, Wikipedia Text Classification

1 Introduction

Text classification is a foundational technology driving many of today's most widely used tools. From search engines that categorize web pages based on relevance, to AI assistants like ChatGPT that analyze and respond to a wide range of queries, text classification is integral to enabling intelligent and context-aware systems. Moreover, businesses rely on these techniques for tasks like spam detection, email filtering, customer feedback analysis, and more. However, most text classification systems are domain-specific, meaning they require significant adjustments or retraining to work across different topics or industries.

The recent advent of pre-trained transformer models, such as BERT, with attention mechanisms has significantly advanced text classification, particularly for complex tasks such as multi-label classification. While these models have shown impressive performance on topic-specific classification tasks, there is still a gap when it comes to generalizing classifiers that can distinguish texts based on their level of expertise—from beginner-level content to expert-level content—independently of the text's specific subject matter.

This paper seeks to design a text expertise classifier capable of distinguishing between texts that are suited for beginner or expert audiences, without being dependent on the specific domain or subject of the text. The proposed study evaluates the performance of statistical features, such as those derived from Zipf's law, contextual embeddings from the USE model, and contextual embeddings from BERT's pre-trained language models. By considering both a general level of expertise and a topic-specific level of expertise, our classifier aims to improve upon existing systems that are often constrained by topic boundaries.

2 Real-World Applications

The ability to classify text expertise on a general level, without being tied to specific topics, has wide-reaching applications. For instance:

- **Educational Tools:** Automatically categorizing content as either beginner or expert-level could help personalize learning materials for students, adapting content to their level of understanding or automating students' essay scoring.
- **Content Curation:** Platforms like news aggregators or knowledge-sharing websites (e.g., Stack Overflow, Wikipedia) could use such a system to recommend articles or tutorials based on a user's expertise level, improving user engagement and experience.
- **Customer Support:** Companies could use expertise classifiers to direct customer queries to appropriate knowledge bases or agents, ensuring that complex issues are handled by experts and simpler queries are routed to beginner-level resources.
- **Job Matching:** Career platforms could benefit by recommending job descriptions or professional articles that match an individual's experience level, improving both candidate engagement and employer satisfaction.

3 Research Objectives and Questions

To address the challenge of creating a topic-independent text expertise classifier, we aim to investigate several approaches based on statistical and contextual models. Specifically, we aim to design a text classifier using the Zipf distribution, the USE model and BERT's language model, considering two dimensions:

1. A general level of expertise (beginner vs. expert).
2. A topic-specific level of expertise (e.g., specialized knowledge within a given field).

This leads us to the following research questions:

1. **Research Question 1:** Can the Zipf distribution be used to identify statistical indicators that differentiate expert-level texts from beginner-level texts?
   We hypothesize that certain statistical properties of word frequencies, captured by the Zipf distribution, may reveal patterns that distinguish complex, expert-level texts from simpler, beginner-level ones.

2. **Research Question 2:** Can we fine-tune BERT's pre-trained model to classify text into beginner or expert categories based on its complexity?
   By leveraging BERT's contextual embeddings, we seek to train a classifier that can understand the nuances of language complexity and classify text according to its level of expertise.

3. **Research Question 3 (Potential):** Is the fine-tuned BERT classifier capable of classifying texts into expert or beginner levels across multiple topics, provided the word distributions of each topic are generated using the Hierarchical Dirichlet Process (HDP)?
   We aim to explore whether BERT can adapt to different topics by using topic-specific word distributions generated through unsupervised learning techniques like HDP.

4. **Research Question 4 (Potential):** Can we extract specific dimensions from BERT's embeddings that encode topic-level expertise, and use these dimensions to classify texts based on their topic-specific expertise level?
   This question explores whether the dense, contextual embeddings generated by BERT can reveal topic-specific levels of expertise that may be useful for fine-grained classification within specialized fields.

4 Related Work

We categorize previous work into two broad classes: non-attention-based text classifiers and attention-based text classifiers.

Non-attention-based classifiers generally fall into two categories: statistical methods [1] and machine learning techniques. Statistical methods, such as those based on word frequency distributions [2] or statistical models like TF-IDF and GloVe, rely on feature extraction and probability calculations, and do not require training. While they perform well on small datasets, they struggle with generalization on more complex data due to their reliance on assumptions (e.g., the independence assumption in GloVe) and their inability to capture nuanced text patterns or semantic relationships. For example, models like TF-IDF represent text as a bag of words (BoW), failing to incorporate contextual meaning or sentence structure [3].

In contrast, machine learning techniques, such as Word2Vec, attempt to address some of these limitations by capturing semantic relationships between words based on their contextual usage. Word2Vec models words that frequently appear in similar contexts as semantically similar in the vector space. However, these methods suffer from unidirectional limitations, as they do not consider the full context of a word's usage in both directions. This issue is partly resolved by bidirectional models such as BiLSTMs, which represent words in a context-sensitive manner, allowing for a richer understanding of word meanings. Statistical feature extraction is often combined with machine learning models to improve classification accuracy, as seen in models like [4], where statistical methods augment the semantic capabilities of machine learning classifiers.

Attention-based text classifiers represent a more recent advance, with the introduction of the attention mechanism enhancing model interpretability and performance. Attention models assign higher importance (i.e., attention scores) to words that are more relevant for classification, such as highly polarized terms in binary classification tasks. For example, researchers investigated how attention scores are assigned to polarized words, showing that these words carry more weight in the model's decision-making process [5].

Recent work on large language models (LLMs) like GPT has also shown promising results in binary text classification tasks [6]. These models, while powerful, often require fine-tuning to achieve optimal performance. The combination of LLMs with statistical features has also been explored to further enhance classification performance. In particular, hybrid or joint models that integrate both contextual understanding from LLMs (such as BERT) and statistical features have gained attention. For example, a joint model approach to classify citizen complaints, combining BERT's contextual embeddings (e.g., [CLS] token) with keyword enhancement derived from word embeddings and TF-IDF scores [7]. This hybrid method improves classification accuracy by leveraging both contextual and statistical information, where the weighted sum of both embeddings is passed to a fully connected neural network (FCNN) for final classification.

5 Data Description

We introduce a dataset consisting of two types of texts: expert-level texts and beginner-level texts. The expert-level texts are sourced from the full

English Wikipedia, while the beginner-level texts are sourced from the Simple English Wikipedia. The dataset contains two columns: text and label. The text column includes the actual text data, and the label column indicates the level of the text. A label of 0 corresponds to beginner-level texts, and a label of 1 corresponds to expert-level texts.

The dataset spans a variety of topic categories, with each category represented in both the expert and beginner-level texts. These categories reflect the breadth of content covered in Wikipedia, and each category is represented in both versions of the text, ensuring a broad and diverse selection of content. Topics may include, but are not limited to, science, history, arts, technology and economy.

The average length of a text in the dataset is 4100 words for a total of 1341 texts.

6 Methodology

6.1 Research Question 1: Zipf's Distribution for Text Classification

The first step of our analysis involves examining the word distribution of our dataset using Zipf's law. We begin by preprocessing the training texts, which includes removing stop words and punctuation, converting all text to lowercase, and applying stemming. Following preprocessing, we tokenize the texts and categorize them into two groups: beginner-level and expert-level texts.

From this analysis, we obtain the slope of the distribution for each class. The beginner-level tokens yield a slope of -1.1586, while expert-level tokens show a slope of -1.2952. This indicates that the expert-level texts exhibit a more pronounced Zipfian distribution, with a higher concentration of high-rank (less frequent) words, while beginner-level texts demonstrate more uniform word usage.

We draw on principles from Zipf's law and the Cumulative Distribution Function (CDF) to classify text documents based on the distribution of word frequencies. The key steps are as follows:

- **Word Frequency Analysis:** Each document's word frequencies are computed across the entire train corpus. Words are ranked based on

how often they appear in the corpus, with the most frequent word assigned rank 1, the second most frequent rank 2, and so on.

- **CDF Calculation:** For each document, a Cumulative Distribution Function (CDF) is calculated using the ranks of the words in the document.

- **Classification:** A threshold is set based on the median CDF of all documents. Documents with a CDF higher than this threshold are classified as belonging to one group (e.g., class 1), and those with a lower CDF are classified into another group (e.g., class 0).

Applied on a test set, this method yields an accuracy of 69% for beginner texts, 81% for expert texts, and an overall accuracy of 73%.

6.2 Research Question 2: Comparison of USE and Fine-tuned BERT Classifiers

- **Baseline: Universal Sentence Encoder (USE)**
  The USE model, developed by Google, generates fixed-length sentence embeddings (512 dimensions) by averaging word embeddings derived from GloVe, followed by processing through a Feedforward Neural Network.

  For classification, we generate embeddings for each training text and apply a mean pooling technique (also known as the centroid technique) to obtain a single average embedding for each class (beginner and expert). These centroid embeddings are then used to classify test texts. To classify a test text, we generate its embedding using USE and compute the cosine similarity between the test embedding and each class centroid. The test text is assigned to the class whose centroid is closest to the test embedding.

  Using this approach, we achieve an accuracy of 76% for beginner texts, 84% for expert texts, and an overall classification accuracy of 79%.

- **Fine-tuned BERT-based Classifier**
  For the second model, we leverage the pre-trained BERT model (bert-base-uncased), which consists of 12 layers, 768 hidden units, and 12 attention heads, totaling 110 million parameters. We fine-tune BERT

to classify texts as either beginner or expert by adding a classification head to the pre-trained model.

We fine-tune BERT end-to-end on our labeled train dataset. The training is performed using CrossEntropyLoss as the loss function and the Adam optimizer for gradient descent.

We employed an independent grid search to identify the optimal dropout rate and learning rate for the model. To prevent overfitting, we incorporated an early stopping strategy during training. Early stopping is a form of regularization that monitors the model's performance on a validation set, and halts training if the validation loss does not improve for a pre-defined number of consecutive epochs. This approach helps ensure that the model does not continue learning unnecessary patterns or noise in the data, thereby improving generalization and preventing overfitting. The grid search and early stopping procedures were critical in fine-tuning the model's hyperparameters and achieving robust performance on unseen data.

After training the model, we extracted the CLS embeddings for each text sample. These embeddings serve as a fixed-length representation of the input texts. For each label (0 or 1), we calculated the centroid embedding by averaging the CLS embeddings of all training samples associated with that label. To classify a test text, we generated its CLS embedding using the trained model and compared it to the centroids of both classes using cosine similarity. The test text was assigned to the class corresponding to the centroid with the highest cosine similarity, enabling the model to classify the test text based on its proximity to the class centroids in the embedding space.

The fine-tuned BERT classifier achieves on the test data an accuracy of 94% for the beginner class, 88% for the expert class, and an overall accuracy of 92%.

7 Results

The following table summarizes the classification accuracies of the different methods evaluated in this study: Zipf Distribution (Threshold-based), USE (Centroid-based), and Fine-tuned BERT.

Method/Approach	Beginner Class Accuracy	Expert Class Accuracy	Overall Accuracy
Zipf Distribution (Threshold-based)	0.69	0.81	0.73
USE (Centroid-based)	0.76	0.84	0.79
BERT (Fine-tuned)	0.94	0.88	0.92

Table 1: Accuracy of Different Classification Methods

8 Conclusion

In this paper, we investigated three approaches for binary text classification aimed at distinguishing texts based on their level of expertise—beginner or expert—using both statistical and machine learning methods. The Zipf distribution provided a statistical lens for understanding word frequency distributions and how they can be used to classify text according to its complexity. The USE model, with its pre-trained embeddings, offered a more contextually aware method for text classification by capturing semantic relationships between text elements. Finally, BERT, with its attention-based architecture, demonstrated the most advanced and accurate approach by leveraging contextual embeddings to understand the nuanced complexity of language and classify texts accordingly.

Our findings suggest that while the Zipf-based method is helpful for certain basic classifications, the USE and BERT-based models provide significantly better performance due to their ability to capture deeper contextual relationships within the text. The results of this study have implications for the development of more accurate and contextually aware text classifiers, particularly in applications that require distinguishing between texts of varying complexity mainly as educational tools. Future work could explore the fine-tuning of these models on more diverse datasets, further improving their performance across a wider range of domains and topics.

9 Open Questions

In light of the promising results from analyzing statistical features and contextual embeddings for expertise classification, several open questions remain that could lead to future research avenues:

1. **Can we develop a hybrid model that integrates statistical features derived from Zipf's law with the contextual knowledge learned by BERT to create a more refined and accurate text**

classifier?
This question explores whether combining the simplicity and mathematical properties of statistical models like Zipf with the deep contextual understanding of BERT could produce a model that leverages the strengths of both approaches, resulting in better classification accuracy.

2. **Can we design a system that uses the semantic dimensions captured by BERT's embeddings to identify topic-specific expertise within multi-topic texts, distinguishing between beginner and expert sections?**
This question focuses on whether BERT's fine-grained representations of language can be leveraged to detect expertise levels within specific topics, even in texts that cover multiple domains. By analyzing these embeddings, we could identify which sections of a multi-topic text are aimed at novice readers and which are intended for experts.

10 References

References

[1] "Text Classification Techniques: A Literature Review," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 13, pp. 117–135, 2018, doi: https://doi.org/10.28945/4066.

[2] Moradi Dakhel, Arghavan, et al., "Assessing Developer Expertise from the Statistical Distribution of Programming Syntax Patterns," *Evaluation and Assessment in Software Engineering*, 21 June 2021, pp. 90–99, https://dl.acm.org/doi/pdf/10.1145/3463274.3463343.

[3] JRKC Jayakody, et al., "Empirical Analysis for the Selection of Baseline Performances for Short Text Classification," vol. 350, 25 Aug. 2023, pp. 335–340, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=arnumber=10253582.

[4] Ramesh, Dadi, and Suresh Kumar Sanampudi, "An Automated Essay Scoring Systems: A Systematic Literature Review," *Artificial Intelligence Review*, vol. 55, 23 Sept. 2021, pp. 2495–2527, https://doi.org/10.1007/s10462-021-10068-2.

[5] Sun, Xiaobing, and Wei Lu, "Understanding Attention for Text Classification," *Association for Computational Linguistics*, 2020.

[6] Abburi, Harika, et al., "Generative AI Text Classification Using Ensemble LLM Approaches," *ArXiv.org*, 2023, https://arxiv.org/abs/2309.07755.

[7] Wang, Yuanhang, et al., "A Joint Attention Enhancement Network for Text Classification Applied to Citizen Complaint Reporting," *Applied Intelligence*, vol. 53, no. 16, 1 Mar. 2023, pp. 19255–19265, https://doi.org/10.1007/s10489-023-04490-y.